

# Advanced practise combining datasets in Excel

Version: 1.0



# Learning intentions

We will be learning **more about how to combine datasets in Excel**, specifically

- How to join datasets when the **key columns have different names**
- How to join datasets with **multiple key columns**
- How to solve problems by **selecting the appropriate join type**

# Background

In previous lessons, we have looked at the theory around combining datasets and then practised combining them.

In this lesson we are going to learn more about using **key columns** and how joining datasets allows you **solve data problems**.



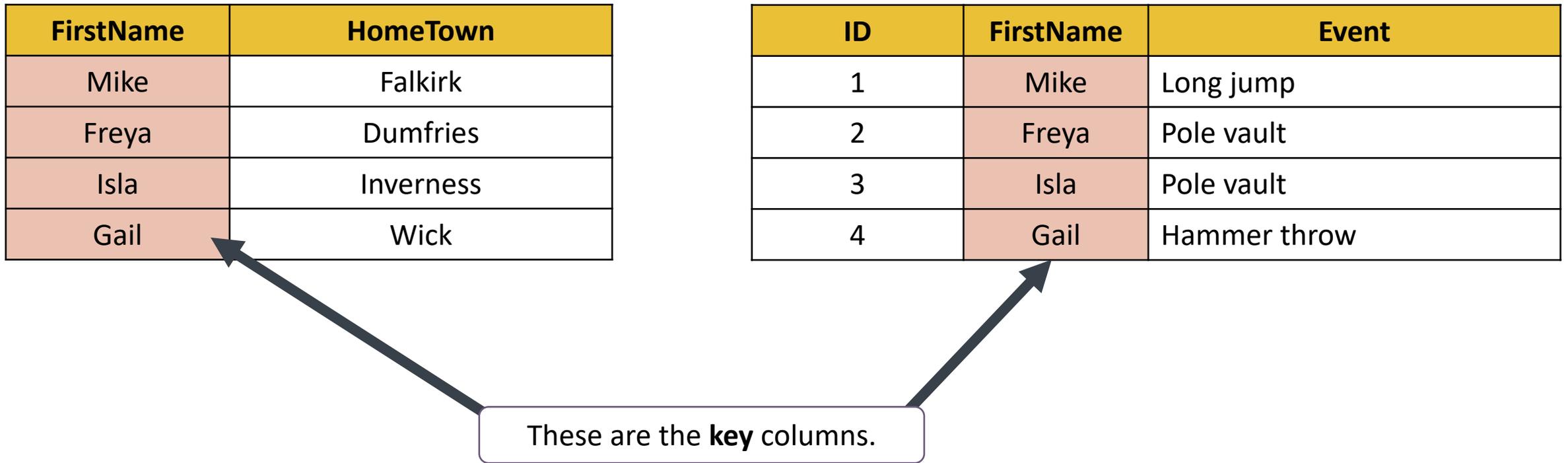
# Reminder... key columns

When joining datasets they need to have at least one column in common. This is called the **key** column and is used to join them.

FirstName	HomeTown
Mike	Falkirk
Freya	Dumfries
Isla	Inverness
Gail	Wick

ID	FirstName	Event
1	Mike	Long jump
2	Freya	Pole vault
3	Isla	Pole vault
4	Gail	Hammer throw

These are the **key** columns.

The diagram illustrates the concept of key columns in a database join. It shows two tables side-by-side. The first table has columns 'FirstName' and 'HomeTown'. The second table has columns 'ID', 'FirstName', and 'Event'. In both tables, the 'FirstName' column is highlighted in a light red color. A callout box at the bottom center contains the text 'These are the key columns.' Two dark grey arrows originate from this box: one points to the 'FirstName' cell in the first row of the first table, and the other points to the 'FirstName' cell in the first row of the second table.

# Key columns with different names

In the last example, the key column in both datasets had the same name – **FirstName**.

Sometimes this isn't the case. For example:

dog_name	max_height
Border Collie	22
Irish Setter	27
Beagle	16
Great Dane	32

animal_name	good_with_children
Border Collie	3
Yorkshire Terrier	5
Chihuahua	1
Beagle	N/A

These are the **key** columns, but they have different names.

# Differently named key columns in Excel

In Excel, you identify the key column in each dataset by clicking on them individually.

Therefore, it **does not matter if the key columns have different names.**

As long as you have identified which column in each dataset you are using as the key column, Excel will still be able to combine the datasets.

Select tables and matching columns to create a merged table.

dog\_name\_height

dog_name	max_height
Border Collie	22
Irish Setter	27
Beagle	16
Great Dane	32

animal\_good\_with\_children

animal_name	good_with_children
Border Collie	3
Yorkshire Terrier	5
Chihuahua	1
Beagle	N/A

# Your turn....



If you were to combine these datasets, which column from each dataset would you use as the key column?

BandName	Genre
Runrig	Celtic Rock
Simple Minds	Pop Rock
Travis	Indie Rock
Big Country	Alternative Rock
The Fratellis	Indie Rock

Album	Artist	YearReleased
The Cutter and the Clan	Runrig	1987
New Gold Dream	Simple Minds	1977
The Man Who	Travis	1990
Costello Music	The Fratellis	2006

# Your turn....



If you were to combine these datasets, which column from each dataset would you use as the key column?

BandName	Genre
Runrig	Celtic Rock
Simple Minds	Pop Rock
Travis	Indie Rock
Big Country	Alternative Rock
The Fratellis	Indie Rock

Album	Artist	YearReleased
The Cutter and the Clan	Runrig	1987
New Gold Dream	Simple Minds	1977
The Man Who	Travis	1990
Costello Music	The Fratellis	2006

These are the key columns.

# Reminder...properties of a key column

The values in a key column need to be:

- **Unique** - the value in each row must be unique, which means there are no duplicate key values
- **Not missing or NULL** – there cannot be any empty or blank values

Additionally, keys must be the **same data type** in both the datasets being joined.



# Composite key columns

Sometimes datasets do not have a single column that fulfils all the properties of a key column (**unique, non-null** and **same data type**).

FirstName	LastName	HomeTown
Mary	Jones	Falkirk
Mary	Smith	Dumfries
Joe	Smith	Dumfries

However, **you can use more than one column** that, when used together, allow you fulfil all the properties of a key columns.

# Definition



## **Composite key**

A combination of two or more columns that can be used to uniquely identify each row in a dataset

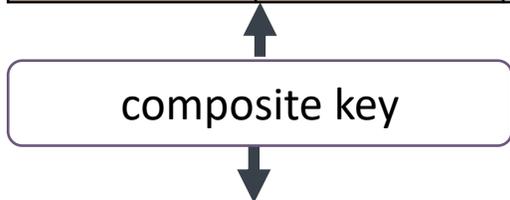
# Show me...



In the datasets below, there isn't a single column that can be used as a key.

However, by using both the **FirstName** and **LastName** columns as the key, the datasets can be joined.

FirstName	LastName	HomeTown
Mary	Jones	Falkirk
Mary	Smith	Dumfries
Joe	Smith	Dumfries



FirstName	LastName	Event
Mary	Jones	High jump
Mary	Smith	Triple jump
Joe	Smith	Long jump

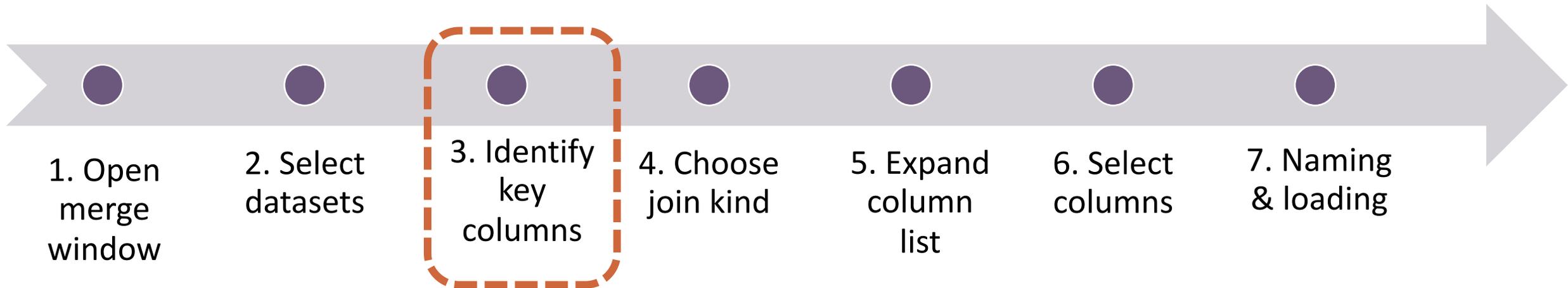


FirstName	LastName	HomeTown	Event
Mary	Jones	Falkirk	High jump
Mary	Smith	Dumfries	Triple jump
Joe	Smith	Dumfries	Long jump

**FirstName** and **LastName** uniquely identify each row in the left and right datasets. Together they form a **composite key**.

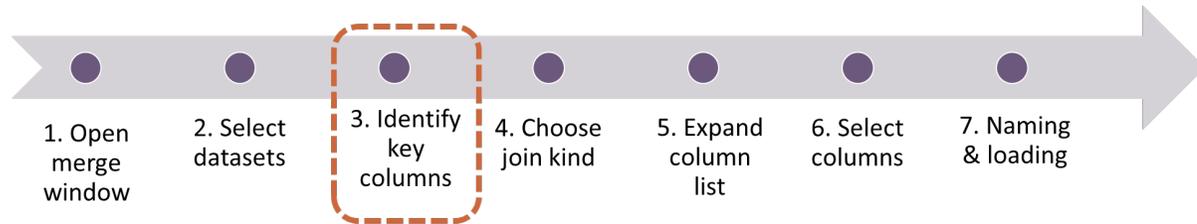
# Reminder: how to join datasets in Excel

Below is a reminder of stages you need to follow when joining datasets in Excel.



When joining using a **composite key** to join datasets in Excel you can follow the same steps, with a slight change to **stage 3**.

# How to join on composite columns in Excel



To join on composite key columns in Excel, when you get to stage 3, you need to,

1. Hold down **Ctrl + click** on the **first key column** in the top dataset
2. Hold down **Ctrl + click** on the **second key column** in the top dataset

Merge

Select tables and matching columns to create a merged table.

population\_under16

town	gender	under_16
Edinburgh	female	36343
Edinburgh	male	38248
Glasgow	female	48895
Glasgow	male	50972
Dundee	female	11638

population\_over16

town	gender	16-64	65_over
Edinburgh	female	179904	42806

Ctrl + click on both columns

# How to join on composite columns in Excel

You now need to repeat **Ctrl + click** on the columns for the bottom dataset.

The **order you click the columns is important.**

You need to make sure that the columns that match have the same numbers against them.

*Hint: Don't worry if you click the columns in the wrong order, you can click them again while holding down the Ctrl button to undo the selection.*

population_under16				
town	1	gender	2	under_16
Edinburgh		female		36343
Edinburgh		male		38248
Glasgow		female		48895
Glasgow		male		50972
Dundee		female		11638

population_over16					
town	1	gender	2	16-64	65_over
Edinburgh		female		179904	42806
Edinburgh		male		175379	33840

Ctrl + click on columns

Next steps

Complete **questions 1 to 4**  
in **section 1** of the  
'Advanced practise combining datasets' workbook.

# How to choose the correct join type to use

When you approach your data and analysis you should always have the problem you are trying to solve in mind.

Data analysts manipulate datasets as they are **not able to solve the problem** they are investigating with the data in form it has been collected in.

By thinking about the problem you are trying to solve will **help decide what type of join** you need to use with the datasets.



*The data science lifecycle framework*

# Joining data to solve problems

You have been asked to find out the name of the country that has the happiest citizens.

You have **two datasets**, one contains a **measure of happiness** and the other the **country name**, so they need to be joined together to answer the specific question.

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
...	...	...

**Dataset name:** country\_profile

**Number of rows:** 156

**Number of columns:** 3

**Duplicate rows:** No

**Missing/blank data items:** No

code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

**Dataset name:** country\_code

**Number of rows:** 244

**Number of columns:** 2

**Duplicate rows:** No

**Missing/blank data items:** No

# Joining data to solve problems

We are now going to work through the steps you need to follow to find out the name of the country that has the happiest citizens....

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
...	...	...

code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
...	...

Combine the datasets

code	happy_score	country_name
LB	5.197	Lebanon
SG	6.262	Singapore
DO	5.425	Dominican Republic
FR	6.592	France
YE	3.380	Yemen
...	...	...

# Your turn....



If we are trying to find out the name of the country with the happiest citizens, which **columns** from each of these dataset do we need in our final dataset?

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
IE	7.021	5.03
...	...	...

*country\_profile*

code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

*country\_code*

# Your turn....



If we are trying to find out the name of the country with the happiest citizens, which **columns** from each of these dataset do we need in our final dataset?

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
IE	7.021	5.03
...	...	...

*country\_profile*

code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

*country\_code*

We will need **code**, **country\_name** and **happy\_score**.

# Labelling the datasets

Before joining datasets, you need to decide which one will be the left or right dataset, as this will impact on which type of join you will use.

The datasets below have been identified as...

## LEFT DATASET

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
IE	7.021	5.03
...	...	...

*country\_profile*

## RIGHT DATASET

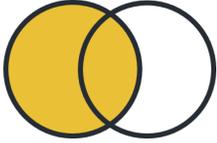
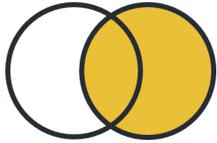
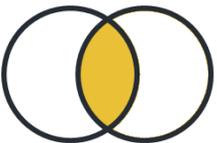
code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

*country\_code*

# Your turn....



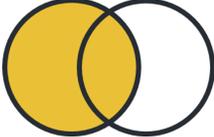
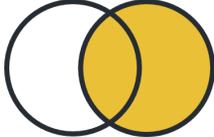
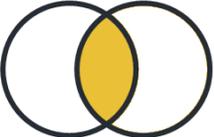
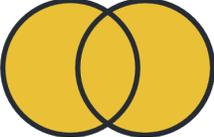
Now we know which columns we want in the final dataset, **what type of join should we use?**

Join Type	Definition
Left Join 	Returns all the values from the left dataset and any matching records from the right dataset
Right Join 	Returns all the values from the right dataset and any matching records from the left dataset
Inner join 	Returns data items whenever there are matching values in both datasets
Outer (full) join 	No information is lost, since it merges any data in either dataset

# Your turn....



We should use a **left join**.

Join Type	Use	Why?
Left Join 	<b>Yes</b>	We want to <b>keep all the information in the left dataset.</b>
Right Join 	<b>No</b>	We don't need the country name if they don't have a happiness score.
Inner join 	<b>No</b>	We don't want to loose a country from the left dataset just because we can't match it in the right dataset.
Outer (full) join 	<b>No</b>	We don't need the names of countries that don't have a score.

# Reminder... joining checklist

Before joining dataset, you need to complete your joining checklist.

Have you identified the **key** column(s)?

---

Are the data items in the key column(s) in the **same data type**?

---

Have you checked for **duplicate rows**?

---

Do you know how **many rows** you expect in the final dataset?

---

Do you know how **many columns** you expect in the final dataset?

---

Do you expect **gaps/missing data items** in your final dataset?

# Your turn...



The first question on the checklist is, “Have you identified the **key** column(s)?”

What are the key columns in the datasets we need to join?

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
IE	7.021	5.03
...	...	...

*country\_profile*

code	country_name
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

*country\_code*

# Your turn...



What are the key columns in the dataset we need to join?

code	happy_score	population_million
MA	5.208	37.08
NZ	7.307	5.12
ES	6.354	47.42
IE	7.021	5.03
...	...	...

code	country
AE	United Arab Emirates
AF	Afghanistan
AL	Albania
AM	Armenia
...	...

**code** is the key column

# Filled in joining checklist

The checklist below has now been filled in for the datasets we need to join,

<input checked="" type="checkbox"/> Have you identified the <b>key</b> column(s)?	Yes
<input checked="" type="checkbox"/> Are the data items in the key column(s) in the <b>same data type</b> ?	Yes
<input checked="" type="checkbox"/> Have you checked for <b>duplicate rows</b> ?	There are no duplicates
<input checked="" type="checkbox"/> Do you know how <b>many rows</b> you expect in the final dataset?	156 (# rows in left dataset)
<input checked="" type="checkbox"/> Do you know how <b>many columns</b> you expect in the final dataset?	3
<input checked="" type="checkbox"/> Do you expect <b>gaps/missing data items</b> in your final dataset?	Might have missing values in the country_name column

# Joining best practice



After joining datasets, you should check that the final dataset looks as you expect.

This includes **checking the number of rows and columns** and whether you have any **missing values** or **duplicate** rows.

# Your turn...



The final dataset has been created by left joining the country\_profile and country\_code datasets.

Which of the following processes do you need to complete now to **find out the name of the happiest country?**

- a) **Plot the data** on a graph?
- b) **Calculate the average** happiness score?
- c) **Sort the dataset** from largest happy\_score to smallest?
- d) **Filter** the dataset on country\_name?

code	happy_score	country_name
LB	5.197	Lebanon
SG	6.262	Singapore
DO	5.425	Dominican Republic
FR	6.592	France
YE	3.38	Yemen
SV	6.253	El Salvador
TD	4.35	Chad
MT	6.726	Malta
...	...	...

# Your turn...



Which of the following processes do you need to complete now to **find out the name of the happiest country?**

- a) Plot the data on a graph?
- b) Calculate the average happiness score?
- c) **Sort the dataset** from largest happy\_score to smallest?
- d) Filter the dataset on country\_name?

**Finland has the happiest citizens.**

code	happy_score ↑	country_name
FI	7.769	Finland
DK	7.600	Denmark
NO	7.554	Norway
IS	7.494	Iceland
NL	7.488	Netherlands
CH	7.480	Switzerland
SE	7.343	Sweden
NZ	7.307	New Zealand
...	...	...

# Joining datasets to solve problems

Here is a summary of the stages we have worked through to join the datasets.



1. What is the **problem** I am trying to answer?

---



2. Which **datasets** can you use to solve the problem?

---



3. Which **columns** do I need in my final dataset?

---



4. Which **type of join** do I need to use?

---



5. Have I completed the **pre-joining checklist**?

---



6. After joining, does my dataset **look as I expect**?

## Next steps

Complete **questions 1 to 7**  
in **section 2**, and **questions 1 and 2**  
in **section 3** of the  
'Advanced practise combining datasets' workbook.

# Learning checklist

I can *describe* what it means to join using a composite key.

I can *join* two simple datasets using Power Query in Excel using a composite key.

I can *choose* the correct join type needed to solve a problem.

# How you can use this lesson



You are free to:

- **Share** – copy and redistribute the material in any medium or format
- **Adapt** – remix, transform and build upon the material

Under the following terms:

- **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for [commercial purposes](#).
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

© 2024. This work is licensed under a [CC BY-NC-SA 4.0 license](#).

Created by effini in partnership with The Data Lab.



# Alternative format

If you require this document in an alternative format, such as large print or a coloured background, please contact

[hello@effini.com](mailto:hello@effini.com)

or

4th Floor, The Bayes Centre  
47 Potterrow  
Edinburgh  
EH8 9BT

